MOLECULAR ECOLOGY

Why the COI barcode should be the community DNA metabarcode for the Metazoa

Journal:	*Molecular Ecology*
Manuscript ID	MEC-18-0597.R1
Manuscript Type:	Opinion
Date Submitted by the Author:	21-Jul-2018
Complete List of Authors:	Andujar, Carmelo; Instituto de Productos Naturales y Agrobiologia, Island Ecology and Evolution Arribas, Paula; Instituto de Productos Naturales y Agrobiologia, Island Ecology and Evolution Yu, Douglas; Kunming Institute of Zoology, ECEC; University of East Anglia, BIO Vogler, Alfried; Imperial College/Natural History Museum of London, Natural Sciences Emerson, Brent; IPNA-CSIC, Ecology and Evolution;
Keywords:	Metabarcoding, barcoding, eDNA, Next Generation Sequencing (NGS), High Throughput Sequencing (HTS)

OPINION ARTICLE

Why the COI barcode should be the community DNA metabarcode for the Metazoa

Carmelo Andújar[1*], Paula Arribas[1], Douglas W. Yu[2,3,4], Alfried P. Vogler[5,6] Brent C. Emerson[1]

1. Grupo de Ecología y Evolución en Islas, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), San Cristóbal de la Laguna, Spain
2. State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China
3. School of Biological Sciences, University of East Anglia, Norwich, Norfolk, UK
4. Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming Yunnan, 650223 China
5. Department of Life Sciences, Natural History Museum, London, UK
6. Department of Life Sciences, Imperial College London, Ascot, UK

* Corresponding author: Carmelo Andújar. Email: candujar@um.es

Key words: Metabarcoding, barcoding, eDNA, Next Generation Sequencing (NGS), High Throughput Sequencing (HTS)

Running Title: COI barcode for metazoan metabarcoding

Abstract

Metabarcoding of complex metazoan communities is increasingly being used to measure biodiversity in terrestrial, freshwater, and marine ecosystems, revolutionizing our ability to observe patterns and infer processes regarding the origin and conservation of biodiversity. A fundamentally important question is which genetic marker to amplify, and although the mitochondrial cytochrome oxidase subunit I (COI) gene is one of the more widely used markers in metabarcoding for the Metazoa, doubts have recently been raised about its suitability. We argue that (i) the extensive coverage of reference-sequence databases for COI, (ii) the variation it presents, (iii) the comparative advantages for denoising protein coding genes, and (iv) recent advances in DNA sequencing protocols argue in favour of standardising for the use of COI for metazoan community samples. We also highlight where research efforts should focus to maximise the utility of metabarcoding.

Introduction

Metabarcoding (Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012; Yu et al., 2012), i.e. the bulk DNA amplification and high-throughput sequencing (HTS) of biological samples, is now a well-established tool for the study of biodiversity, as reflected by the rapid growth in the number of published studies since the early applications to bacteria and fungi (e.g., Buée et al., 2009; Hamady, Walker, Harris, Gold, & Knight, 2008) (Fig.1). Metabarcoding has been applied to DNA from diverse biological sources using a wide range of laboratory procedures and addressing manifold questions about spatial and temporal biodiversity patterns (e.g., Deiner et al., 2017; Taberlet, Bonin, Zinger, & Coissac, 2018). The most straightforward application of metabarcoding is the acquisition of DNA data from bulk

specimen samples. These are mixed species assemblages that have been extracted from their habitat matrix and combined for a single DNA extraction, followed by PCR amplification with 'universal' primers. This approach, referred to as community DNA metabarcoding (cMBC) (Deiner et al., 2017) is increasingly being applied to biodiversity inventories and biomonitoring in marine (e.g., Fonseca et al., 2010; Leray & Knowlton, 2015), terrestrial (e.g., Arribas, Andújar, Hopkins, Shepherd, & Vogler, 2016; Ji et al., 2013) and freshwater environments (e.g., Andújar et al., 2018; Elbrecht & Leese, 2017) (See Fig. 1). Although there are technical differences, metabarcoding of metazoan communities can also be conducted on DNA extractions directly from the external medium, such as soil or water, to gather 'environmental DNA' (eDNA; see glossary) (Taberlet, Coissac, Hajibabaei, & Rieseberg, 2012; Deiner et al., 2017 for a comparison between community and environmental DNA metabarcoding)

A key design consideration for metazoan metabarcoding is the selection of the DNA marker to be amplified, a choice that may greatly affect the number of species and taxonomic groups detected and the accuracy of species identifications against marker-specific reference databases. Taxonomic bias associated with PCR primer choice has been the main reason to question the utility of several markers for DNA metabarcoding (Deagle et al., 2014; Taberlet, Coissac, Pompanon, et al., 2012), including the mitochondrial cytochrome oxidase 1 gene (COI or *cox1*) where is located the standard barcode region (COI-bcr) for metazoan DNA taxonomy (Hebert, Cywinska, Ball, & DeWaard, 2003; also see the Consortium for the Barcode of Life, CBOL; http:// www.barcodeoflife.org/). Additional considerations for fragment choice in metazoan metabarcoding are the state of preservation of the DNA template (eDNA is often fragmented; e.g., Deagle, Eveson, & Jarman, 2006), read-length limitations of widely-used parallel-sequencing methods (e.g, a maximum read length of 300 bp of the

Illumina technology, limiting paired-sequencing to amplicons of maximally ≈450 bp; e.g., Fadrosh et al., 2014), and potential co-amplification of concomitant microbial DNA (e.g., Stat et al., 2017). Due to these concerns, marker choice for metazoan metabarcoding lacks a universally agreed approach, which has resulted in a proliferation of primers with different taxon specificities and degree of universality.

The above-mentioned concerns are well-founded in the case of eDNA metabarcoding (Deagle et al., 2014), where DNA is often poorly preserved and frequently includes high proportions of microbial DNA (e.g., Stat et al., 2017; Yang et al., 2014). However, concerns regarding DNA integrity and co-amplification of microbial DNA are largely inconsequential for cMBC. It is largely for reasons of presumed taxonomic bias for PCR amplification of the COI-bcr that many studies have abandoned this locus, in favour of primers matching highly conserved binding sites with a presumed more even coverage of all taxa present. The most widely used alternatives are the nuclear ribosomal genes coding for the small subunit (SSU or 18S rRNA) (Capra et al., 2016; Creer et al., 2010), the large subunit (LSU or 28S rRNA) (Hirai, Kuriyama, Ichikawa, Hidaka, & Tsuda, 2014), the internal transcribed spacer 2 (ITS2) (Anslan & Tedersoo, 2015; Avramenko et al., 2017), and the mitochondrial small [*rrnS* or 12S rRNA] (Machida, Kweskin, & Knowlton, 2012) and large subunit rRNA [*rrnL* or 16S rRNA] (Elbrecht et al., 2016; Saitoh et al., 2016). The lack of consensus over the choice of metabarcode markers, even within the same target community, carries the risk of poor standardisation and low comparability among studies, which ultimately hampers the development of an efficient, universal system for biodiversity discovery and monitoring using cMBC.

Here we argue in favour of the COI-bcr as a standard for bulk-sampled metazoan cMBC and support our position with four sets of arguments. We revisit two points that have

made the COI-bcr the fragment of choice for barcoding in metazoans and equally apply to cMBC: the availability of large COI-bcr reference databases, and the level of nucleotide variation of COI-bcr that is appropriate for the taxonomic assignment of amplicons at the species level. Our third point is that sequencing errors and spurious sequence assemblies can be robustly identified by bioinformatic processing based on the predicted variation in protein coding regions and the limited length variation in COI-bcr. Finally, recent evidence regarding potential taxonomic amplification bias associated with the COI-bcr, a key reason for questions about its utility, can be overcome by improved design of primers. We conclude by focussing on the benefits and synergies that can emerge from standardisation, and provide recommendations for future research and applications.

1. Large COI-bcr reference databases provide a powerful link to taxonomic identity

The utility of a reference sequence database for metabarcoding is a function of: (i) the inherent power of the marker for taxonomic assignment; (ii) the taxonomic coverage (number of species and phylogenetic diversity represented in the database) and depth (number of individuals sequenced per species) of reference sequences, and (iii) the adequate formatting and curation of the database and its accessibility to taxonomic-assignment software packages. The taxonomic coverage and depth of COI-bcr is unparalleled. Public records at the BOLD online database (Ratnasingham & Hebert, 2007) include 1,240,301 sequences of >500 bp in length, representing 102,254 species (accessed 26 May 2018). Taking into account sequences on BOLD that are yet to be made publically available, there are 5,542,839 sequences of which 3,150,643 are identified to species representing 191,568 animal species.

COI-bcr resources clearly exceed those available for any other DNA marker for animals. For example, *rrnL* and *rrnS* include 256,372 and 137,603 sequences on GenBank (Benson et al., 2014), while SSU include 149,119 sequences (searches on 26 May 2018 at GenBank for sequences of >500 bp within Metazoa). There were 135,416 and 127,065 metazoan sequences for LSU and SSU, respectively, on the SILVA database (Quast et al., 2013) (searches on 26 May 2018). Additionally, Machida et al. (2017) have recently constructed the *Midori* database, which includes all mitochondrial genes of the Metazoa, including GenBank records available prior to September 2015. *Midori* also provides a quantitative measure of the available taxonomic coverage of different mtDNA gene regions, demonstrating the dominant representation of COI-bcr (583,043 sequences) which greatly exceeds the next-most represented regions of cytochrome oxidase b (*cob*; 223,247 sequences) and *rrnL* (146,164 sequences), and is represented for more species in almost all animal phyla (Machida et al., 2017).

As a reference database increases is size, the probability of false taxonomic assignment is reduced and placement to lower taxonomic ranks is improved (Somervuo et al., 2016). In this context, it is worth noting the expected future growth of the COI-bcr reference dataset due to ongoing geographically and taxonomically focused campaigns. When such campaigns incorporate historic type specimens into barcode projects (e.g., Hausmann et al., 2016), stronger linkage is forged between traditional taxonomic systems and reference sequences. Barcode campaigns that employ rigorous taxonomic identification of voucher specimens also provide a necessary step forward to identify database sequences that have been incorrectly assigned taxonomically, as it has been shown to occur in the Genbank database (Mioduchowska, Jan, Gołdyn, Kur, & Sell, 2018).

In addition to the availability of reference sequences, tools are needed to manage such large databases and facilitate taxonomic classification of the unprecedented volume of sequences obtained by metabarcoding (Somervuo et al., 2016). The BOLD website itself was not designed for the large-volume searches needed by metabarcoding, although an application programming interface (v4.boldsystems.org/index.php/api_home, accessed 8 Mar 2018) allows automated queries via the R *bold* package (github.com/ropensci/bold, accessed 8 Mar 2018), and a new BOLD database interface, suitable for large-volume queries, has recently been made publically available (mbrave.net, accessed 8 Mar 2018). Additionally, the *Midori* web server (www.reference-midori.info, accessed 8 Mar 2018) provides three taxonomic-assignment methods (RDP Classifier (Wang, Garrity, Tiedje, & Cole, 2007), SPINGO (Allard, Ryan, Jeffery, & Claesson, 2015), and SINTAX (Edgar, 2016a)) for volume queries.

2. Taxonomic identification and intraspecific structure – two for the price of one

Thanks to its relatively high mutation rate, COI-bcr (and other mitochondrial genes) is a powerful marker to detect intraspecific variation, which can be separated from interspecific variation using various algorithms for sequence clustering and phylogenetic rates (e.g., Hebert & Gregory, 2005; Pons et al., 2006; Puillandre, Lambert, Brouillet, & Achaz, 2012; J. Zhang, Kapli, Pavlidis, & Stamatakis, 2013) and thus improves the ability to distinguish closely related and cryptic species (Candek & Kuntner, 2015). In contrast, the *SSU* gene, widely used to characterise marine meiofauna and soil fauna (Capra et al., 2016; Creer et al., 2010; Yang et al., 2014) has a comparatively lower mutation rate, increasing the probability that related species may share the same sequence (Andújar et al., 2018; Tang et al., 2012). As well as

compromising species identification, such limited variation will also underestimate both alpha and beta diversity, fundamental metrics for meaningful ecological conclusions from metabarcode studies.

The high mutation rate of COI-bcr and resulting intraspecific variation have been widely used to investigate the structuring of genetic variation below the species level (e.g., Bucklin, Steinke, & Blanco-Bercial, 2011; Goodall-Copestake, Tarling, & Murphy, 2012) and to inform about ecological and evolutionary processes at the community level (e.g., Baselga et al., 2013; Emerson et al., 2017). HTS data have not taken advantage of this property of the COI-bcr, largely because sequence quality has been perceived to be low, and it is effectively removed as sequence variants are clustered into OTUs. However, as read quality improves, simple clustering can be replaced by direct use of HTS reads, albeit after stringent denoising that removes spurious sequence variants (Callahan, McMurdie, & Holmes, 2017; Edgar, 2016b). Denoising can be particularly efficient for COI-bcr due to the predictable pattern of nucleotide variation within protein-coding mitochondrial genes and the almost complete absence of length variation within the COI-bcr (see below). Indeed, recent work by Elbrecht, Vamos, Steinke, & Leese (2018) demonstrates the ability to recover intraspecific genetic variation from cMBC data, opening the door for the simultaneous analysis of species diversity and intraspecific variation for cMBC at the whole-community or even ecosystem level.

3. The advantage of protein-coding genes to identify spurious sequences

Bioinformatic steps for removing non-target sequences that can originate from PCR errors, sequencing errors, amplification of pseudogenes, and chimeric rearrangements (Edgar, 2016b; Schirmer et al., 2015) can be carried out more robustly for protein-coding genes compared to

188 ribosomal gene regions (Ramirez-Gonzalez et al., 2013; Ranwez, 2011). This is due to the
189 pattern of variation of protein-coding mitochondrial genes, where: (i) some amino-acid
190 residues are highly conserved; (ii) nucleotide variation is biased toward the third base
191 positions of codons; and (iii) indels are almost completely absent (Ramirez-Gonzalez et al.,
192 2013). Thus, COI-bcr metabarcode reads leading to stop codons or indels are clear targets for
193 removal, and denoising can also take advantage of known patterns of variation in protein
194 coding sequences to detect (i) atypical ratios of synonymous/nonsynonymous mutations, (ii)
195 atypical amino acid changes compared to representative consensus sequences, and (iii)
196 atypical distributions of variation with respect to codon position (Ramirez-Gonzalez et al.,
197 2013; Ranwez, 2011). These features can potentially be integrated in the denoising process to
198 retain only well supported genetic variants from COI-bcr HTS reads.

199

200 4. Comprehensive and informative surveys with better design of primers

201 Metabarcoding using fragments within the COI-bcr has been associated with the incomplete
202 recovery of species from mock communities ('dropouts') (e.g., Clarke, Soubrier, Weyrich, &
203 Cooper, 2014; Yu et al., 2012), and as a consequence the utility of the COI-bcr has been
204 questioned (Deagle et al., 2014). A key reason for dropouts is high heterogeneity in primer
205 binding sites and thus differential PCR efficiencies across variable templates, which results in
206 taxonomic bias during PCR amplification. A related consequence is that differences in
207 amplification efficiency complicate the use read frequencies as proxy measures of species
208 abundance or biomass (Krehenwinkel et al., 2017; Piñol, Mir, Gomez-Polo, & Agustí, 2015).
209 Proposed remedies include the use of multiple, taxon-specific primers on the same sample
210 (Drummond et al., 2015; Stat et al., 2017).

Despite earlier concerns (Deagle et al., 2014), the extent to which the COI-bcr produces taxonomic bias in metazoan cMBC is unclear. Performance varies among studies, with many factors potentially explaining variation, such as target taxa, relative abundance and body size, specimen preservation, laboratory procedures, primers choice, and PCR conditions. For example, the low recovery of species documented in some studies (Brandon-Mong et al., 2015; Clarke et al., 2014; Elbrecht et al., 2016) also coincides with the use of mostly non-degenerate primers (Table 1). Yu et al. (2012) used degenerate LCO1490 and HC02198 primers and inherently low-coverage 454 pyrosequencing to achieve promising results for cMBC, recovering up to 76% of the species from mock pools of known composition, including 12 different orders within the Arthropoda. Although a dropout of 24% is undesirable, Yu et al. (2012) showed that even this level of dropout did not prevent metabarcoding data from providing correct estimates of community-level metrics, namely alpha and beta diversity, and thus metabarcoding data were reliable inputs to decision-making (Ji et al., 2013).

Studies using redesigned, degenerate primers for various subregions of the COI-bcr have continued to reduce dropout in cMBC of Metazoa (Andújar et al., 2018; Arribas et al., 2016; Beng et al., 2016; Elbrecht & Leese, 2017; Leray et al., 2013; Prosser, Velarde-Aguilar, León-Règagnon, & Hebert, 2013; Saitoh et al., 2016) (Table 1). In a study of aquatic taxa including 52 macroinvertebrates, Elbrecht & Leese (2017) showed that the use of degenerate primers within the COI-bcr recovered almost all input taxa (42/42 insects; 9/10 other taxa) and resulted in improved estimation of relative abundances, a result that outperformed even the *rrnL* primer set (41/42 species of Insecta and 2/10 other taxa). However, it should be noted that the estimation of species abundance from metabarcode data is controversial and requires further research, probably requiring calibration studies using known amounts of

DNA (Bista et al., 2018; Krehenwinkel et al., 2017; Thomas, Deagle, Eveson, Harsch, & Trites, 2016). In another study of whole-community freshwater invertebrates (Andújar et al., 2018), cMBC with SSU universal primers and degenerate COI-bcr primers resulted in the detection of 2-4 times higher number of 97%-similarity OTUs (operational taxonomic units) with COI-bcr, including the main insect orders inhabiting freshwater ecosystems (Diptera, Coleoptera, Ephemeroptera, and Trichoptera), plus Crustacea, Rotifera, and Annelida (Andújar et al., 2018). However, amplification of nematodes and platyhelminthes was poor, and requires different primer sets (e.g., Prosser et al., 2013). With increasing knowledge of taxon-specific problems, primer design and combinations of primer sets can be adapted to generate increasingly complete community inventories and improved species abundance data.

Concluding remarks

We conclude that the much greater number of COI-bcr reference sequences, the broader taxonomic coverage and resolution of these sequences, combined with recent improvements in COI-bcr primer design, argue for the COI-bcr region as the marker of choice cMBC of bulk metazoan samples. An important caveat here is that we do not include eDNA samples in our recommendation. In the case of eDNA, the target region for the Metazoa is frequently present only at very low concentrations compared to microbial DNA (Stat et al., 2017), and it is widely found, although not generally published, that most primers within the COI-bcr amplify large proportions of microbial species (e.g., Yang et al., 2014). This fact remains the strongest reason for the use of mitochondrial rRNA markers that are much less affected by this type of cross-amplification. Ultimately, with the increasing availability of whole mitochondrial genomes, MBC studies using COI-bcr and other markers can be linked (Arribas et al. 2016).

Looking forward, we identify the following key areas of research and development for cMBC: (1) Continued and increased funding for alpha taxonomy, DNA barcoding campaigns, and the development and maintenance of the BOLD database, increasing its functionality regarding metabarcode data. Regarding other public databases (e.g., GenBank), effort is required to identify sequences with incorrect taxonomic assignment to avoid their use as reference data (Mioduchowska et al., 2018). (2) Development and validation of detailed and standardisable methods for field work and extraction of the target specimens from their habitat matrix (water, soil, sediment etc) (e.g., Arribas et al., 2016; Fonseca et al., 2010). (3) Continued design and validation of primers for DNA fragments within the COI-bcr, with the aim of standardizing fragments of choice within the COI-bcr to maximise comparability among studies. For example, the Leray-Geller primer set (Leray et al., 2013) is now widely used because the amplicon length of 313 bp matches the read lengths of paired-end Illumina sequencing, but this primer set was largely designed for marine organisms, and thus could probably be improved upon for terrestrial taxa. Other promising primer sets include those used by Elbrecht & Leese (2017) for a fragment of 316 bp (BF1-BR2) and Shokralla et al. (2014) and Andújar et al., (2018) for a fragment around 400 bp (pair of primers Ill_B_F-Ill_B_R and Ill_B_F-Fol-degen-rev respectively). A related issue is that various primers target different, and frequently non-overlapping regions of the COI-bcr, which limits the direct comparsions among metabarcoding studies, in particular for those taxa without exact matches to sequences in the reference database. (4) Development and validation of denoising methods for the recovery of intraspecific genetic variation from cMBC data. This will include evolutionary models of sequence variation that go beyond the current error models based on prevalent technical artifacts of the sequencing procedure (e.g. Schirmer et al., 2015) or read abundances (Edgar, 2016b). (5) Continued development, validation, and improved availability

of methods for taxonomic assignment (e.g., Somervuo et al., 2016; A. Zhang, Hao, Yang, & Shi, 2016).

Much progress has been made in the field of cMBC in recent years, and the potential for cMBC as an integrated tool for biodiversity monitoring and management is clearly recognised (e.g. Bush et al, 2017). Standardising for the COI-bcr for cMBC and focussing on the above suggestions should increase the reliability of metabarcode data for management, policy and decision-making, while also facilitating greater comparability across independent studies.

Figure 1

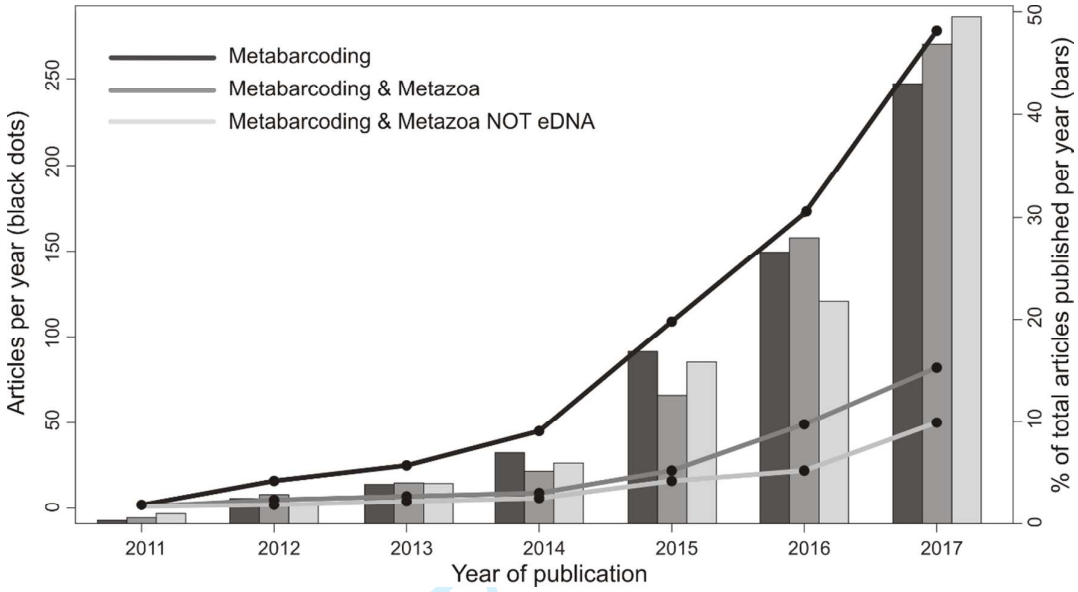

Figure 1 Temporal evolution of scientific publications on the topic *metabarcoding* (Black line; TS = *metabarcoding*); *metabarcoding on metazoans* (Dark grey line: TS=(metabarcoding) NOT TS =(*micro* OR *bacteria* OR *myco* OR *archaea* OR fungi OR plant); and *metabarcoding on metazoans excluding eDNA studies* (Light grey line: TS=(metabarcoding) NOT TS =(*micro* OR *bacteria* OR *myco* OR *archaea* OR fungi OR plant OR eDNA OR environmental DNA). Black dots: number of publications per year for each search. Bars: proportion of the total publications of each search per year. Searches were performed on the Web of Science (23-04-2018), including the Science Citation Index Expanded, Social Science Citation Index, Arts and Humanities Citation Index, and Conference Proceedings Citation Index–Science databases for all years and restricted to article types "article" and "review".

Table 1. Overview of studies providing data on the performance of different fragments within the COI-bcr on community DNA metabarcoding (cMBC) for Metazoa.

Reference	Target taxa	Type of primers	Amplicon length(bp)	vitro/ silico	Results
(Prosser et al., 2013)	Nematoda	Degenerate	650	vitro	89.5% (85/95) sequencing success on diverse parasitic nematode lineages, including members of three orders and eight families.
(Beng et al., 2016)	Arthropoda	Degenerate	ca. 400	vitro	100% in-vitro PCR efficiency on a wide range of arthropods (Chilopoda, Araneae, Hymenoptera, Blattodea, Mantodea, Coleoptera, Orthoptera, Lepidoptera, and Hemiptera)
(Beng et al., 2016)	Arthropoda	Degenerate	ca. 400	silico	100% detection succes after in silico sequencing six mock communities with known arthropod composition (37 ref sequences from Genbank)
(Arribas et al., 2016)	Acari and Collembola	Degenerate	650	vitro	Detection of >100 species of Acari and Collembola from 28 families. Recovery against 79 barcoded voucher specimens in the same samples was 95% (75/79)
(Andújar et al., 2018)	Freshwater invertebrates	Degenerate	420*	vitro	COI outperformed SSU except for Nematodes and Platyhelminthes
(Saitoh et al., 2016)	Collembola	Degenerate	314	vitro	100% (7/7) recovery in mock communities. In complex soil samples, cMBC on COI outperformed morphology, and provided a similar recovery to *rrnL* (16S).
(Yu et al., 2012)	Arthropoda	Degenerate	650	vitro	Recovery rates of 76% for already barcoded species by Sanger.
(Elbrecht et al., 2016)	Freshwater invertebrates	Non-degenerate	650	vitro	Recovery of 90% (38/42) insects and 50% (5/10) other taxa in a mock community.
(Elbrecht & Leese, 2017)	Freshwater invertebrates	Degenerate	316**	vitro	Recovery of 100% (42/42) insects and 90% (9/10) other taxa in a mock community.
(Clarke et al., 2014)	Insects	Non- or low-degenerate	Several pairs	silico	For every pair of primer, recovery of <75% of insect species with complete mitochondrial genome available. *rrnL*(16S) recovered >90%.
(Clarke et al., 2014)	Insects	Non- or low-degenerate	Several pairs	vitro	Recovery of the same or less taxa than with *rrnL* (16S) on a mock community of 14 taxa.
(Brandon-Mong et al., 2015)	Arthropoda	Only forward degenerate	313	vitro	Recovery of 91% (71/78) species on a mock community with representatives for Aranea, Blattodea, Coleoptera, Diptera, Hemiptera, Hymenptera, Lepidoptera, Matodea, Odonata, Orthoptera and Collembola
(Krehenwinkel et al., 2017)	Arthropoda	Degenerate	313 and 418	vitro	Recovery of 95% (41/43) on a mock community including 19 orders in the Arachnida, Crustacea, Hexapoda & Myriapoda. Same or higher recovery than other fragments tested (Cytb, 12s, 18s, 28s, H3).

* Refers to primers Ill_B_F and Fol-degen-rev. ** Refers to primers BF1 and BR2. *** Refers to primers mlColintF and HCO2198

BOX 1 Glossary

DNA barcoding. Method for the taxonomic identification of specimens based on the sequencing of diagnostic DNA sequence regions. It was first proposed by Hebert et al (2003). Frequently used barcodes (i.e., DNA fragments used for DNA barcoding) include the COI gene for Metazoa, rbcL for plants, ITS for fungi and *rrnL* (16s) for bacteria.

High-throughput sequencing (HTS). Techniques that allow the simultaneous sequencing of millions of DNA fragments.

DNA metabarcoding. DNA amplification and high-throughput sequencing of a DNA extract derived from a biological sample composed of a mix of DNA from different source species, each represented by one or more individuals. After bioinformatic procedures for quality filtering, resulting DNA sequences can be subject to molecular identification using barcode reference databases.

Environmental DNA (eDNA) metabarcoding. DNA metabarcoding targeting DNA directly isolated from environmental samples such as soil, sediments or water, among others (Taberlet, Coissac, Hajibabaei, & Rieseberg, 2012). DNA sources contributing to eDNA include the breakdown of body parts from organisms together with faeces, mucus, skin cells, organelles, gametes or even extracellular DNA.

Community DNA metabarcoding (cMBC). DNA metabarcoding targeting DNA isolated from bulk mixtures of specimens that have been extracted from their habitat matrix.

Invertebrate ingested DNA (iDNA) metabarcoding. DNA metabarcoding targeting vertebrate genetic material that is extracted from invertebrates (such as leeches, mosquitoes, or ticks, among others). Can be considered as an particular case of eDNA metabarcoding, as the DNA sources are of ingested material or faeces.

Degenerate primer. Mixture of DNA oligonucleotides that differ in base composition for one or several nucleotide positions (degenerate positions). In practice, it means that different variants of a particular oligo are synthesized and mixed to be used as primers on a PCR reaction. The higher the proportion of degenerate positions, the more degenerate a primer is.

Universal primers. PCR primers, degenerate or not, with the potential to amplify a particular DNA fragment within a broad taxonomic scope (e.g. all Metazoa, all Arthropoda, all Crustacea, etc). Although full universality (i.e. amplifying all species within the taxonomic scope) is unlikely, primers are often referred to as universal when they broadly function across the phylogenetic diversity within a given taxonomic scope.

REFERENCES

Allard, G., Ryan, F. J., Jeffery, I. B., & Claesson, M. J. (2015). SPINGO: A rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics*, *16*(1), 1–8. doi:10.1186/s12859-015-0747-1

Andújar, C., Arribas, P., Gray, C., Bruce, C., Woodward, G., Yu, D. W., & Vogler, A. P. (2018). Metabarcoding of freshwater invertebrates to detect the effects of a pesticide spill. *Molecular Ecology*, *27*(1), 146–166. doi:10.1111/mec.14410

Anslan, S., & Tedersoo, L. (2015). Performance of cytochrome c oxidase subunit I (COI), ribosomal DNA Large Subunit (LSU) and Internal Transcribed Spacer 2 (ITS2) in DNA barcoding of Collembola. *European Journal of Soil Biology*, *69*, 1–7. doi:10.1016/j.ejsobi.2015.04.001

Arribas, P., Andújar, C., Hopkins, K., Shepherd, M., & Vogler, A. P. (2016). Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods in Ecology and Evolution*, *7*(9), 1071–1081. doi:10.1111/2041-210X.12557

Avramenko, R. W., Redman, E. M., Lewis, R., Bichuette, M. A., Palmeira, B. M., Yazwinski, T. A., & Gilleard, J. S. (2017). The use of nemabiome metabarcoding to explore gastro-intestinal nematode species diversity and anthelmintic treatment effectiveness in beef calves. *International Journal for Parasitology*, *47*(13), 893–902. doi:10.1016/j.ijpara.2017.06.006

Baselga, A., Fujisawa, T., Crampton-Platt, A., Bergsten, J., Foster, P. G., Monaghan, M. T., & Vogler, A. P. (2013). Whole-community DNA barcoding reveals a spatio-temporal continuum of biodiversity at species and genetic levels. *Nature Communications*, *4*(May), 1892, DOI: 10.1038/ncomms2881. doi:10.1038/ncomms2881

Beng, K. C., Tomlinson, K. W., Shen, X. H., Surget-Groba, Y., Hughes, A. C., Corlett, R. T., & Slik, J. W. F. (2016). The utility of DNA metabarcoding for studying the response of arthropod diversity and composition to land-use change in the tropics. *Scientific Reports*, *6*(September). doi:10.1038/srep24965

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2014). GenBank. *Nucleic Acids Research*, *42*(D1). doi:10.1093/nar/gkt1030

Bista, I., Carvalho, G. R., Tang, M., Walsh, K., Zhou, X., Hajibabaei, M., … Creer, S. (2018). Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Molecular Ecology Resources*, 0–2. doi:10.1111/1755-0998.12888

Brandon-Mong, G.-J., Gan, H.-M., Sing, K.-W., Lee, P.-S., Lim, P.-E., & Wilson, J.-J. (2015). DNA metabarcoding of insects and allies: an evaluation of primers and pipelines. *Bulletin of Entomological Research*, *105*(06), 717–727. doi:10.1017/S0007485315000681

Bucklin, A., Steinke, D., & Blanco-Bercial, L. (2011). DNA Barcoding of Marine Metazoa. *Annual Review of Marine Science*, *3*(1), 471–508. doi:10.1146/annurev-marine-120308-080950

Buée, M., Reich, M., Murat, C., Morin, E., Nilsson, R. H., Uroz, S., & Martin, F. (2009). 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *The New Phytologist*, *184*(2), 449–56. doi:10.1111/j.1469-8137.2009.03003.x

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*, *11*(12), 2639–2643. doi:10.1038/ismej.2017.119

Candek, K., & Kuntner, M. (2015). DNA barcoding gap: Reliable species identification over morphological and geographical scales. *Molecular Ecology Resources*, *15*(2), 268–277. doi:10.1111/1755-0998.12304

Capra, E., Giannico, R., Montagna, M., Turri, F., Cremonesi, P., Strozzi, F., … Pizzi, F. (2016). A new primer set for DNA metabarcoding of soil Metazoa. *European Journal of Soil Biology*, *77*, 53–59. doi:10.1016/j.ejsobi.2016.10.005

Clarke, L. J., Soubrier, J., Weyrich, L. S., & Cooper, A. (2014). Environmental metabarcodes for insects: In silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, *14*(6), 1160–1170. doi:10.1111/1755-0998.12265

Creer, S., Fonseca, V. G., Porazinska, D. L., Giblin-Davis, R. M., Sung, W., Power, D. M., … Thomas, W. K. (2010). Ultrasequencing of the meiofaunal biosphere: Practice, pitfalls and promises. *Molecular Ecology*, *19*(SUPPL. 1), 4–20. doi:10.1111/j.1365-294X.2009.04473.x

Deagle, B. E., Eveson, J. P., & Jarman, S. N. (2006). Quantification of damage in DNA recovered from highly degraded samples--a case study on DNA in faeces. *Frontiers in Zoology*, *3*, 11. doi:10.1186/1742-9994-3-11

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., Taberlet, P., Taberlet, P., … Hajibabaei, M. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, *10*(9), 1789–1793. doi:10.1098/rsbl.2014.0562

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., … Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895. doi:10.1111/mec.14350

Drummond, A. J., Newcomb, R. D., Buckley, T. R., Xie, D., Dopheide, A., Potter, B. C. M., … Nelson, N. (2015). Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience*, *4*(1). doi:10.1186/s13742-015-0086-1

Edgar, R. (2016a). SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *BioRxiv*, 074161. doi:10.1101/074161

Edgar, R. (2016b). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*, 081257. doi:10.1101/081257

Elbrecht, V., & Leese, F. (2017). Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment. *Frontiers in Environmental Science*, *5*(April), 1–11. doi:10.3389/fenvs.2017.00011

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J.-N., …

Leese, F. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ*, *4*, e1966. doi:10.7717/peerj.1966

Elbrecht, V., Vamos, E. E., Steinke, D., & Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA metabarcoding data, 1–13. doi:10.7287/peerj.preprints.3269v3

Emerson, B. C., Casquet, J., López, H., Cardoso, P., Borges, P. A. V., Mollaret, N., … Thébaud, C. (2017). A combined field survey and molecular identification protocol for comparing forest arthropod biodiversity across spatial scales. *Molecular Ecology Resources*, *17*(4), 694–707. doi:10.1111/1755-0998.12617

Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., & Ravel, J. (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, *2*(1), 6. doi:10.1186/2049-2618-2-6

Fonseca, V. G., Carvalho, G. R., Sung, W., Johnson, H. F., Power, D. M., Neill, S. P., … Creer, S. (2010). Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*, *1*(7), 98. doi:10.1038/ncomms1095

Goodall-Copestake, W. P., Tarling, G. A., & Murphy, E. J. (2012). On the comparison of population-level estimates of haplotype and nucleotide diversity: A case study using the gene cox1 in animals. *Heredity*, *109*(1), 50–56. doi:10.1038/hdy.2012.12

Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J., & Knight, R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods*, *5*(3), 235–237. doi:10.1038/nmeth.1184

Hausmann, A., Miller, S. E., Holloway, J. D., deWaard, J. R., Pollock, D., Prosser, S. W. J., & Hebert, P. D. N. (2016). Calibrating the taxonomy of a megadiverse insect family: 3000 DNA barcodes from geometrid type specimens (Lepidoptera, Geometridae). *Genome*, *59*(9), 671–684. doi:10.1139/gen-2015-0197

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B*, *270*(1512), 313–21. doi:10.1098/rspb.2002.2218

Hebert, P. D. N., & Gregory, T. R. (2005). The promise of DNA barcoding for taxonomy. *Systematic Biology*, *54*(5), 852–859. doi:10.1080/10635150500354886

Hirai, J., Kuriyama, M., Ichikawa, T., Hidaka, K., & Tsuda, A. (2014). A metagenetic approach for revealing community structure of marine planktonic copepods. *Molecular Ecology Resources*, *15*(1), 68–80. doi:10.1111/1755-0998.12294

Ji, Y., Ashton, L., Pedley, S. S. M., Edwards, D. D. P., Tang, Y., Nakamura, A., … Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, *16*(10), 1245–57. doi:10.1111/ele.12162

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, *7*(1). doi:10.1038/s41598-017-17333-x

Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized

samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences*, *2014*(July), 201424997. doi:10.1073/pnas.1424997112

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*(1), 1–14. doi:10.1186/1742-9994-10-34

Machida, R. J., Kweskin, M., & Knowlton, N. (2012). PCR primers for metazoan mitochondrial 12S ribosomal DNA sequences. *PLoS ONE*, *7*(4), 1–6. doi:10.1371/journal.pone.0035887

Machida, R. J., Leray, M., Ho, S. L., & Knowlton, N. (2017). Data Descriptor: Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Scientific Data*, *4*(January), 1–7. doi:10.1038/sdata.2017.27

Mioduchowska, M., Jan, M., Gołdyn, B., Kur, J., & Sell, J. (2018). Instances of erroneous DNA barcoding of metazoan invertebrates: Are universal cox1 gene primers too "universal"? *Plos One*, 1–16. doi:10.1371/journal.pone.0199609

Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, *15*(4), 819–830. doi:10.1111/1755-0998.12355

Pons, J., Barraclough, T., Gomez-Zurita, J., Cardoso, A., Duran, D., Hazell, S., … Vogler, A. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, *55*(4), 595–609. doi:10.1080/10635150600852011

Prosser, S. W. J., Velarde-Aguilar, M. G., León-Règagnon, V., & Hebert, P. D. N. (2013). Advancing nematode barcoding: A primer cocktail for the cytochrome c oxidase subunit I gene from vertebrate parasitic nematodes. *Molecular Ecology Resources*, *13*(6), 1108–1115. doi:10.1111/1755-0998.12082

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, *21*(8), 1864–1877. doi:10.1111/j.1365-294X.2011.05239.x

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., … Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), 590–596. doi:10.1093/nar/gks1219

Ramirez-Gonzalez, R., Yu, D. W., Bruce, C., Heavens, D., Caccamo, M., & Emerson, B. C. (2013). PyroClean: denoising pyrosequences from protein-coding amplicons for the recovery of interspecific and intraspecific genetic variation. *PloS One*, *8*(3), e57615. doi:10.1371/journal.pone.0057615

Ranwez, V. (2011). MACSE : Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons, *6*(9). doi:10.1371/journal.pone.0022594

Ratnasingham, S., & Hebert, P. D. N. (2007). BARCODING, BOLD : The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes*, *7*(April 2016), 355–

364. doi:10.1111/j.1471-8286.2006.01678.x

Saitoh, S., Aoyama, H., Fujii, S., Sunagawa, H., Nagahama, H., Akutsu, M., … Nakamori, T. (2016). A quantitative protocol for DNA metabarcoding of springtails (Collembola). *Genome*, *59*(9), 705–723. doi:10.1139/gen-2015-0228

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, *43*(6), e37. doi:10.1093/nar/gku1341

Shokralla, S., Gibson, J. F., Nikbakht, H., Janzen, D. H., Hallwachs, W., & Hajibabaei, M. (2014). Next-generation DNA barcoding: Using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, *14*(5), 892–901. doi:10.1111/1755-0998.12236

Somervuo, P., Yu, D., Xu, C., Ji, Y., Hultman, J., Wirta, H., & Ovaskainen, O. (2016). Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *BioRxiv*. doi:10.1101/070573

Stat, M., Huggett, M. J., Bernasconi, R., Dibattista, J. D., Newman, S. J., Harvey, E. S., … Tina, E. (2017). Ecosystem biomonitoring with eDNA : metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports*, (September), 1–11. doi:10.1038/s41598-017-12501-5

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA for biodiversity research and monitoring*. Oxford, UK: Oxford University Press.

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. *Molecular Ecology*, *21*(8), 1789–93. doi:10.1111/j.1365-294X.2012.05542.x

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–50. doi:10.1111/j.1365-294X.2012.05470.x

Tang, C. Q., Leasi, F., Obertegger, U., Kieneke, a., Barraclough, T. G., & Fontaneto, D. (2012). The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences*, *109*(40), 16208–16212. doi:10.1073/pnas.1209160109

Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H., & Trites, A. W. (2016). Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources*, *16*(3), 714–726. doi:10.1111/1755-0998.12490

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. doi:10.1128/AEM.00062-07

Yang, C., Wang, X., Miller, J. A., de Blécourt, M., Ji, Y., Yang, C., … Yu, D. W. (2014). Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. *Ecological Indicators*, *46*, 379–389. doi:10.1016/j.ecolind.2014.06.028

Yu, D., Ji, Y., Emerson, B., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity

soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, *3*(4), 613–623. doi:10.1111/j.2041-210X.2012.00198.x

Zhang, A., Hao, M., Yang, C., & Shi, Z. (2016). BarcodingR: an integrated R package for species identification using DNA barcodes. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.12682

Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics (Oxford, England)*, *29*(22), 2869–76. doi:10.1093/bioinformatics/btt499

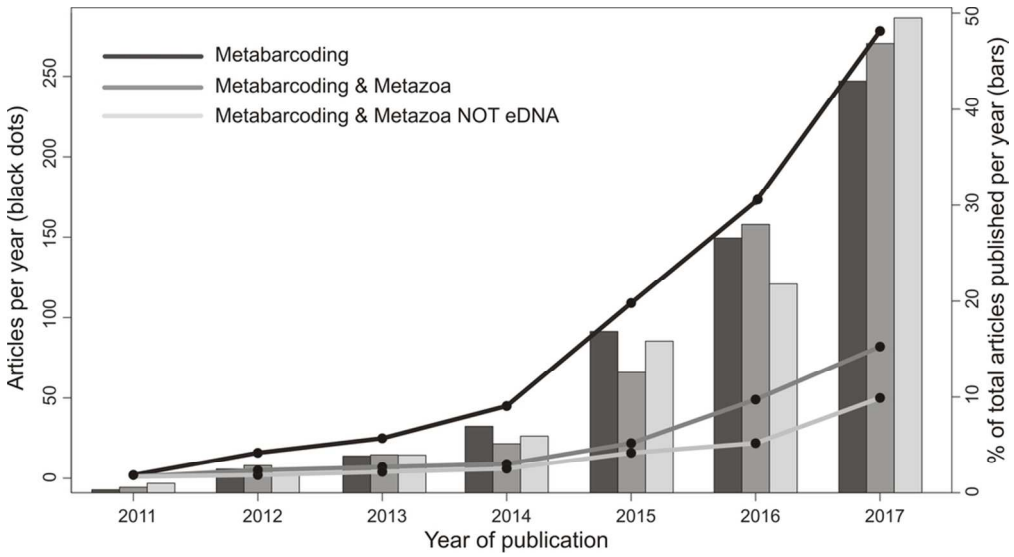

Figure 1 Temporal evolution of scientific publications on the topic metabarcoding

91x50mm (300 x 300 DPI)